Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification

Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification

Sayaka Shiota[1], *Fernando Villavicencio*[2], *Junichi Yamagishi*[2], *Nobutaka Ono*[2], *Isao Echizen*[2], *Tomoko Matsui*[3]

[1]Tokyo Metropolitan University, Hino, Tokyo, 191-0065, Japan.
[2]National Institute of Informatics, Chiyoda, Tokyo, 101-8430, Japan.
[3]The Institue of Statistical and Mathematics, Tachikawa, 190-8562, Japan.

Abstract

This paper proposes a novel countermeasure framework to detect spoofing attacks to reduce the vulnerability of automatic speaker verification (ASV) systems. Recently, ASV systems have reached equivalent performances equivalent to those of other biometric modalities. However, spoofing techniques against these systems have also progressed drastically. Experimentation using advanced speech synthesis and voice conversion techniques has showed unacceptable false acceptance rates and several new countermeasure algorithms have been explored to detect spoofing materials accurately. However, the countermeasures proposed so far are based on the acoustic differences between natural speech signals and artificial speech signals, expected to become gradually smaller in the near future. In this paper, we focus on voice liveness detection, which aims to validate whether the presented speech signals originated from a live human. We use the phenomenon of pop noise, which is a distortion that happens when human breath reaches a microphone, as liveness evidence. This paper proposes pop noise detection algorithms and shows through an experimental study that they can be used to discriminate live voice signals from artificial ones generated by means of speech synthesis techniques.
Index Terms: automatic speaker verification, voice liveness detection, anti-spoofing, countermeasure, pop noise

1. Introduction

It is well known that biometric authentication has an important role in reliable management systems nowadays [1, 2]. Automatic speaker verification (ASV) is also an easy-to-use biometric authentication system using only speakers' voice samples. Recently, the performance of the ASV techniques has been improved as a result of e.g. i-Vector [3, 4] or PLDA (probabilistic linear discriminant analysis) [5] developments, and there are a lot of reports regarding state-of-the-art schemes that show potential to support mass-market adoption. Meanwhile, speech synthesis [6, 7] and speech transformation [8], which are technologies to generate natural-sounding artificial speech with the targeted speaker's voice from a given text or an inputted speech waveform uttered by someone else, have progressed. They are also active and important research topics in speech information processing because the technologies may help individuals with vocal or communicative disabilities for instance. However, such technologies can be used to falsify profiles or identities and perform spoofing attacks against ASV systems, representing a serious challenge to the successful operation of these systems [9–11]. Research on the definition and development of countermeasures for the detection of spoofing attacks already exists [12–15]. Conventionally, attacks of three different natures are considered: replay, speech synthesis, and voice conversion. Countermeasure strategies are mainly based on comparing acoustic features of artificial signals with those of natural ones [16–18]. Spectral, F0 and modulation-related features are among the features used to compute the countermeasures. However, we expect the acoustic differences between artificial and natural speech to gradually become smaller and eventually marginal in the near future.

Looking at other biometrics fields, we see that face, fingerprint, and even iris recognition systems also suffer from spoofing attacks, and researchers have continued to develop several countermeasures to overcome this problem [19–21]. One of the most effective countermeasures in other biometrics fields is to use a "liveness detection" framework that ensures that the person attempting authentication is alive. For image processing fields, it has been reported that liveness detection frameworks have reduced vulnerability significantly [22–24]. We can use the same concept for the ASV system and propose a countermeasure algorithm based on voice liveness detection (VLD), so we can detect spoofing materials more robustly. An important question is how we ensure the liveness of presented speech signals to validate whether the presented signals are originated from a live human or not. For this purpose, in this paper we focus on pop noise detection. Since pop noise is a common distortion in speech occurring when human breath reaches a microphone and is poorly reproduced by loudspeakers [25, 26], it seems reasonable to consider it as natural evidence of liveness at the input of an authentication system. A measure that takes into account the presence of pop noise phenomena might therefore represent therefore a good basis to discriminate between *live* or *played* speech (via loudspeakers). This paper proposes two VLD detection strategies to reduce the vulnerability of ASV systems. To evaluate the effectiveness of the proposed VLD frameworks, we have recorded some speech on a small database including voice samples with pop noise. An experimental evaluation was carried out to explore the performance of the proposed techniques, showing, as it will be furthermore reported, significant benefits when incorporating them as VLD modules within the ASV process.

The outline of this paper is as follows. In section 2, the framework of the voice liveness detection is showed, and pop noise extraction algorithms are illustrated in section 3. Section 4 describes design of database that includes pop noise. Section 5 and section 6 presents the experimental results and conclusions.

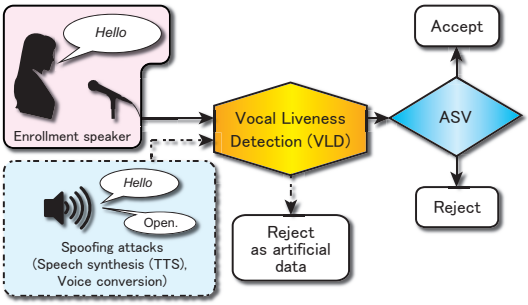

Figure 1: Overview of automatic speaker verification system including VLD module

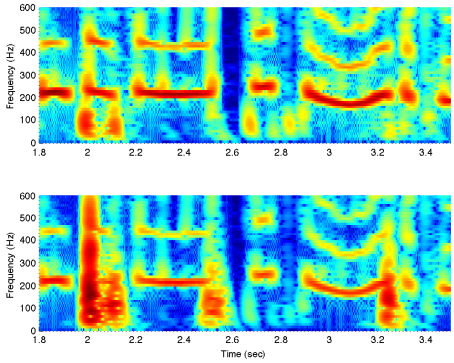

Figure 2: Spectrogram comparison of recording using (top) or not using pop filter (bottom). Significant differences can be seen at low frequency at locations perceived to have pop noise.

2. Voice liveness detection for speaker verification system

2.1. Attack to speaker verification systems

The potential for ASV to be spoofed is well recognized and there is growing interest in assessing the vulnerabilities of ASV systems and developing countermeasures [9, 10]. The countermeasures target three main types of spoofing attacks: replay, speech synthesis, and voice conversion. Each type of attack is defined as follows:

- Replay: replay of pre-recorded utterances of the target speaker.
- Speech synthesis: automatic generation of synthesized speech signals of the targeted speaker based on any input text.
- Voice conversion: conversion of attacker's natural voice towards that of the targeted speaker.

Several countermeasures against each type of spoofing attack have been reported. We can simply use text-prompted ASVs and change prompts every time to protect against replay attacks [27, 28]. However, no methods have reached a fundamental solution against the spoofing attacks using speech synthesis and voice conversion. Considering the actual procedures for spoofing attacks, all spoofing attacks have to play spoofing speech via loudspeakers. In other words, if we can distinguish speech produced by a live human from speech played via loudspeakers, we can protect against all types of spoofing attacks including even attacks using unknown voice conversion and speech synthesis methods.

2.2. Framework of voice liveness detection

Figure 1 shows a diagram of an automatic speaker verification system including the VLD module. The VLD module aims to reject all speech signals that do not include liveness evidence regardless of spoofing type. Speaker verification is conducted as usual in a subsequent module. Although this figure illustrates a sequential combination of VLD and ASV modules, it is also possible to carry out the VLD and ASV modules simultaneously.

What is the liveness evidence included in a speech waveform? The VLD needs to detect and capture characteristics included only in speech produced by a live human. The human voice can be briefly described as a result of acoustic shaping in the vocal tract of the airflow produced following the interaction between various elements such as lungs, vocal chords, and lips. Then, to record the sound, the resulting airflow is transformed to an acoustical signal when it is captured via a microphone. As a consequence of spontaneous strong breathing the convolution process between the airflow and the vocal cavities may result in a sort of perceived plosive burst, commonly know as pop noise, which can be captured via a microphone. On the other hand, the acoustic conditions change when this same sound is played via loudspeakers, commonly resulting in a poor reproduction of pop noise phenomena. Thus, by detecting pop noise events, we may be able to distinguish live human voices from playback sound via loudspeakers.

3. Voice liveness detection algorithms

To capture the phenomenon of pop noise as liveness evidence, this paper proposes two VLD detection strategies to reduce the vulnerability of ASV systems.

3.1. Low-frequency-based single channel detection

Pop noise in single channel signals are found in speech waveforms as sudden irregular modulations of strong energy within varying durations typically ranging between 20 and 100 msec. This phenomenon appears as high energy regions at very low frequency compared to when using a pop noise filter, as shown in Figure 2. This gives us a clue to define a simple strategy for detection.

More precisely, we firstly define the measure $LF_{nrg}(k)$ as the average of the Fourier transform (FT) bins within the interval $[0, LF_{max}]$. The frequency precision should be high enough to explore a very low-band with more than a single FT bin. Accordingly, an analysis window of size N, corresponding to a precision of $10Hz$ in the FT and $LF_{max} = 40Hz$ was found as a sufficient choices. Note that LF_{max} might be set below expected pitch values in order to not consider energy contributions from harmonic content. Following, $LF_{nrg}(k)$ was computed over frames of size N with a hop-size of $M = N/8$ and pop noise events $P_{loc}(i)$ were identified as the maxima of $LF_{nrg}(k)$ with values larger than three times its standard-deviation, keeping a minimum distance of $D = 1.5N$ between candidates.

The boundaries are estimated by approximation in the neighbourhood of each $P_{loc}(i)$ according to two conditions: firstly, a drop in $LF_{nrg}(k)$ of $LF_{dr} = 0.35$ times the value at $P_{loc}(i)$. Then, the boundaries are extended if the absolute value of the derivative of $LF_{nrg}(k)$ is higher than $LF_{ddr} = 0.35$ times its value at $P_{loc}(i)$. With these conditions we aim to as-

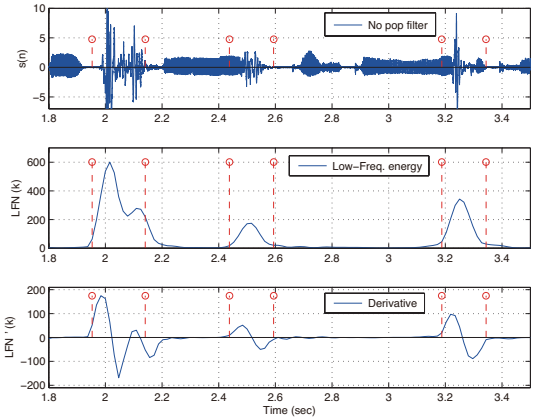

Figure 3: Example of pop noise detection based on single channel method. Time-domain signals (top), average low-band energy (middle), its derivate (bottom), and the detected pop noise boundaries (red dotted).

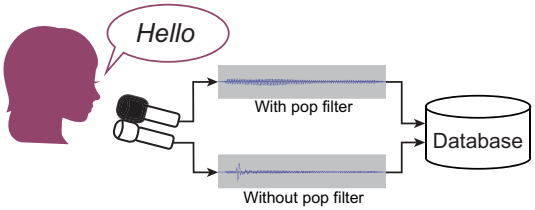

Figure 4: Recording process in two channel method

sert a minimum/maximum energy variation (velocity) once it is ensured: similarly, there will be a relative increment/drop in the pop noise energy. An example can be seen in Figure 3, which shows the waveform of the recording version (top). It also shows the computed $LF_{nrg}(k)$ for the waveform with pop noise (middle) and its derivate (bottom). The detected boundaries are denoted by the red dotted intervals.

Although the configuration of the processing parameters should be manually verified for significant pop noised cases, the suggested parameters, empirically found, showed sufficient performance on samples of several speakers of our database.

3.2. Subtraction-based pop noise detection with two channels

The pop noise detection algorithm using a single channel microphone is focused on low-frequency energy. To capture the whole frequency components of the pop noise, another pop noise detection algorithm is proposed here.

In the second algorithm, two microphones are used and only one of them has a pop filter as shown in Figure 4. Let $F_x(b,w)$ and $F_p(b,w)$ be the short-time Fourier transforms (STFT) of the filtered speech and non-filtered speech respectively, where b and w stand for the indices of time frame and angular frequency. In the two channel method, assuming that only $F_p(b,w)$ includes pop noise, it is estimated by subtracting the ordinary speech component from $F_p(b,w)$ by using $F_x(b,w)$ as follows.

$$D(b,\omega) = F_p(b,\omega) - C(\omega)F_x(b,\omega), \tag{1}$$

where $C(\omega)$ represents a compensation filter between the frequency characteristics of the two channels. An estimate of $C(\omega)$ to minimize $\sum_{b,\omega}|D(b,\omega)|^2$ can be represented as follows.

$$C(\omega) = \frac{\sum_b F_p(b,w)F_x(b,w)^*}{\sum_b |F_x(b,w)|^2}, \tag{2}$$

where * denotes complex conjugate.

4. Design of database

Since the proposed framework focuses on speech signals that include pop noise, a database of speech signals that includes pop noise is required. Recently, the NIST SRE database [29] has been used globally for the evaluation sets of ASV systems. However, the database provides conversational telephone speech and it contains no pop noise, so the proposed framework could not evaluate the conventional databases. Therefore, a new database including pop noise signals is required to be constructed. It is well know that some kinds of microcphones are very sensitive to breath noise [30, 31]. However, there is no preliminary information about pop noise recording and microphone types to be used. Then, we have used three types of microphones as below:

- Microphone with a voice recorder (VOICE) (SONEY ECM-DM5P)
- Compatible microphone with camcorder (CAM) (SONY ECM-XYST1M)
- Microphone with a headset (HEADSET) (SHURE SM10A-CN)

Two microphones of each type are used where one has a pop filter and the other does not (Fig. 4). That is, we designed a six-channel microphone system (Fig. 5). The characteristics of the microphones are as follows. The VOICE microphone is most sensitive to pop noise, and even when using a pop filter, pop noise is often obtained. There is a clear difference between the waveforms of the CAM microphone with a pop filter and those without any pop filter when compared to those of the VOICE microphones. The waveforms of the HEADSET microphones are almost the same with a pop filter and without any pop filter; nevertheless, the HEADSET microphones were set closest to the speaker's mouth. The speech signals were sampled at a 48 kHz with a 16 bit rate.

We have recorded a total of 17 female speakers of Japanese. Each speaker reads out 100 sentences in total. Half of the sentences are know to all the speakers and the other half are randomly selected from Japanese News paper Article Sentences (JNAS) [32], and each speaker uses a different set of randomly selected sentences. The 50 common sentences are chosen based on phonetic coverage. We also pre-selected relatively short sentences from the JNAS corpus before the random selection of the rest of the 50 sentences.

5. Evaluation experiments

5.1. Experimental conditions

We used 30 randomly selected utterances for each microphone without the pop filter for each speaker as live samples of test data. The spoofing materials used in our experiments were constructed based on the statistical parametric speech synthesis framework described in [6]. Its speaker adaptation techniques in this framework allow the generation of a synthetic voice using as little as a few minutes of recorded speech from the target speaker [33]. The speaker adaptation algorithm used is structural variational Bayesian linear regression [34]. We used 50

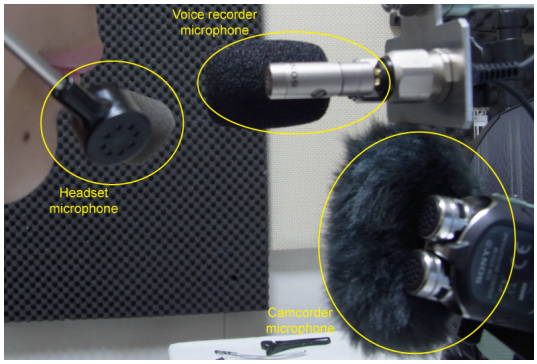

Figure 5: Six microphones are used simultaneously for recording speech data with and without pop filters. Distance from and position in relation to speaker's mouth for each microphone are fixed roughly.

common sentences recorded via the headset microphone with the pop filter to perform the speaker adaptation of speech synthesis systems (because the pop filter is always used for normal recordings of speech synthesis data). Using the speech synthesizers of individual target speakers, we synthesized artificial speech signals for spoofing. The texts used for speech synthesis are the above randomly selected utterances of each speaker. The spoofing materials were then played with a loudspeaker (BOSE 111AD) toward the video camera and condenser microphones.

For the ASV system, we used the standard GMM-UBM-based speaker verification method [35], and the speaker-dependent models of individual speakers in the ASV system were constructed using the 50 common and 20 randomly selected sentences of each speaker recorded via the headset microphone with a pop filter. In this paper, we investigate the effectiveness of the VLD module, and we do not focus on using the state-of-the-art ASV system. The number of mixtures were set to 2048, and the UBM was trained by about 23000 utterances from JNAS database [32], which is the standard speech database for automatic speech and speaker recognition area in Japan. For the STFT analysis, the Hamming window is selected as a window function, and the window width and the window shift are set to 4096 and 2048 points.

5.2. Experimental results

Table 1 shows the equal error rate (EER) of the VLD methods with the test date with a single channel algorithm and two channels algorithms. For each algorithm, the EER is calculated when the percentage of misclassified live voice (false positive rate) is equal to the percentage of misclassified artificial voice (false negative rate). From the results of the single channel algorithm, we were able to be capture, significantly, human liveness information via voice recorder microphone and headset microphone. Even though the performance of the VLD module is dependent on the microphone types, this result illustrates that the VLD framework is effective to reduce the vulnerability of ASV systems. In the single channel method, we assume that the phenomenon of pop noise is strongly affected at low frequency. The two channel method also shows the effectiveness of using pop noise for VLD module. However, comparing the single channel method with the two channel method, the two channel method was obtained small improvement. This means that pop noise affects to the voice at low-frequency. Consequently, the proposed VLD framework is precisely effective method against the spoofing method.

Table 1: EERs of VLD algorithms with some microphones

Microphone	Single ch.	Two ch.
VOICE	4.73%	29.11%
CAM	36.06%	45.52%
HEADSET	**3.95%**	**5.88%**

Table 2: EERs of the ASV system with test data which includes Spoofing Attacks data (w/ SA) or not (w/o SA). And the EERs of the VLD+ASV system.

microphone	w/o SA	w/ SA	VLD+ASV	
			single ch.	two ch.
VOICE	5.49%	5.53%	5.48%	5.49%
CAM	4.69%	6.61%	5.23%	5.30%
HEADSET	4.28%	6.61%	4.45%	4.28%

Secondary, VLD+ASV system which was combined the VLD module with the ASV system was evaluated. The procedure of the VLD+ASV system is shown in Figure 1. Table 2 illustrates the EERs of the ASV system with or without pop noise, and the EER of the VLD+ASV system. From the results, the spoofing attacks degrade the ASV performance. In this experiment, since the spoofing attacks are made by enrollment speech recorded with HEADSET microphone, the spoofing attacks became weak via the VOICE microphone. However, the EERs via HEADSET or CAM microphones is strongly affected from the spoofing attacks. Besides, the EER of the VLD+ASV system is reduced the vulnerabilities to the spoofing attacks adequately. These results indicate that the proposed framework is definitely effective against the spoofing attack sounds.

6. Conclusion

In this paper, novel VLD algorithms are proposed To reduce the vulnerabilities to spoofing attacks using speaker verification systems, the ASV systems have to recognize whether an input voice is live. This quality is known as liveness and ensures that the measured characteristics come from a live human being and are captured at the time of verification. The proposed algorithms focused on capturing the pop noises in an input voice because live humans produce pop noises unconsciously when they speak. One algorithm based on a single microphone was proposed to capture the distortion at low frequency as pop noise. Another algorithm based on two microphones was proposed to capture the pop noises by comparing the filtered channel with the non-filtered channel. To evaluate our proposed algorithms, a database that contains the pop noises was constructed. From the experimental results, we showed that the VLD algorithms could capture the pop noises accurately and hence can discriminate live voice signals from artificial ones. Our future work includes trials using a larger database and extension of the VLD algorithms to the time series settings. We also need to assess the robustness of the proposed method against a waveform concatenation-based synthesis method that uses speech recordings with pop noise or voice conversion where the input speech has pop noises.

7. Acknowledgements

This work was supported in part by Grant-in-Aid for Young Scientists (Start-up), 25880026, and Grant-in-Aid for scientific Research (B), 26280066.

8. References

[1] A. Jain, P. Flynn, and A. Ross, "Handbook of biometrics," 2007.

[2] N. Poh and J. Korczak, "Hybrid biometric person authentication using face and voice features," in *Audio- and Video-Based Biometric Person Authentication*, ser. Lecture Notes in Computer Science, J. Bigun and F. Smeraldi, Eds. Springer Berlin Heidelberg, 2001, vol. 2091, pp. 348–353.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[4] "NIST i-vector Challenge 2014," http://www.nist.gov/itl/iad/mig/ivec.cfm.

[5] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, Oct 2007, pp. 1–8.

[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[7] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, May 1996, pp. 373–376 vol. 1.

[8] Y. Stylianou, "Voice transformation: A survey," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3585–3588.

[9] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Interspeech*, 2013, pp. 925–929.

[10] N. W. D. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, *Speaker recognition anti-spoofing*. Book Chapter in "Handbook of Biometric Anti-spoofing", Springer, S. Marcel, S. Li and M. Nixon, Eds., 2014, June 2014.

[11] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[12] L.-W. Chen, W. Guo, and L.-R. Dai, "Speaker verification against synthetic speech," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, Nov 2010, pp. 309–312.

[13] Z.-Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition." in *INTERSPEECH*, 2012.

[14] M. Faundez-Zanuy, M. Hagmller, and G. Kubin, "Speaker verification security improvement by means of speech watermarking," *Speech Communication*, vol. 48, no. 12, pp. 1608 – 1619, 2006, {NOLISP} 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639306000653

[15] M. Nematollahi, S. Al-Haddad, S. Doraisamy, and M. Ranjbari, "Digital speech watermarking for anti-spoofing attack in speaker recognition," in *Region 10 Symposium, 2014 IEEE*, April 2014, pp. 476–479.

[16] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the i -vector space," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 4, pp. 821–832, April 2015.

[17] R. McClanahan, B. Stewart, and P. De Leon, "Performance of i-vector speaker verification and the detection of synthetic speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3779–3783.

[18] J. Gaka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, no. 0, pp. 143 – 153, 2015.

[19] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG - Proceedings of the International Conference of the*, Sept 2012, pp. 1–7.

[20] D. Yambay, J. Doyle, K. Bowyer, A. Czajka, and S. Schuckers, "Livdet-iris 2013 - iris liveness detection competition 2013," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, Sept 2014, pp. 1–8.

[21] N. Evans, S. Li, S. Marcel, and A. Ross, "Guest editorial special issue on biometric spoofing and countermeasures," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 4, pp. 699–702, April 2015.

[22] B. Toth, "Liveness detection: Iris," in *Encyclopedia of Biometrics*, S. Li and A. Jain, Eds. Springer US, 2009, pp. 931–938.

[23] S. Schuckers, "Liveness detection: Fingerprint," in *Encyclopedia of Biometrics*, S. Li and A. Jain, Eds. Springer US, 2009, pp. 924–931.

[24] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision ECCV 2010*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, vol. 6316, pp. 504–517.

[25] G. Elko, J. Meyer, S. Backer, and J. Peissig, "Electronic pop protection for microphones," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, Oct 2007, pp. 46–49.

[26] Y. Hsu, "Spectrum analysis of base-line-popping noise in mr heads," *Magnetics, IEEE Transactions on*, vol. 31, no. 6, pp. 2636–2638, Nov 1995.

[27] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2, April 1993, pp. 391–394 vol.2.

[28] D. Delacretaz and J. Hennebert, "Text-prompted speaker verification experiments with phoneme specific mlps," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, May 1998, pp. 777–780 vol.2.

[29] "NIST Speaker Recognition Evaluation (SRE)," http://www.itl.nist.gov/iad/mig/tests/spk/.

[30] Y. Nishida, T. Hori, T. Suehiro, and S. Hirai, "Monitoring of breath sound under daily environment by ceiling dome microphone," in *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, vol. 3, 2000, pp. 1822–1829 vol.3.

[31] K. Naraharisetti, "Enhancement of breathing signal using delayless subband adaptive filter with hpf," in *Signal Processing and Information Technology (ISSPIT), 2010 IEEE International Symposium on*, Dec 2010, pp. 177–181.

[32] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[33] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[34] S. Watanabe, A. Nakamura, and B.-H. Juang, "Structural bayesian linear regression for hidden markov models," *Journal of Signal Processing Systems*, vol. 74, no. 3, pp. 341–358, 2014.

[35] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verificaion using adapted gaussian mixture models," *Degital Signal Orocess*, vol. 10, no. 1, pp. 19–41, 2000.